fMRI Data Modelling

Mark W. Woolrich[a]

[a]Oxford University Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB), UK

1. INTRODUCTION

FMRI (Functional Magnetic Resonance Imaging) is a powerful non-invasive tool in the study of the function of the brain, used by psychologists, psychiatrists, and neurologists. FMRI can give high quality visualization of the location of activity in the brain resulting from sensory stimulation or cognitive function. It allows, for example, the study of how the healthy brain functions, how it is affected by different diseases, how it attempts to recover after damage, and how drugs can modulate activity or post-damage recovery.

After an fMRI experiment has been designed and carried out, the resulting data must be passed through various analysis steps before the experimenter can get answers to questions about experimentally-related activations at the individual or multi-subject level. This document gives a brief overview of the most commonly-used analysis pipeline: data pre-processing, temporal linear modelling and activation thresholding. For more details, see Jezzard et al. (2001).

2. FMRI DATA

In a typical fMRI session a low-resolution functional volume is acquired every few seconds. (MR volumes are often also referred to as "images" or "scans"). Example slices from an fMRI volume are shown in Figure 1. Over the course of the experiment, 100 volumes or more are typically recorded. In the simplest possible experiment, some images will be taken whilst stimulation* is applied, and some will be taken with the subject at rest. Because the images are taken using an MR sequence which is sensitive to changes in local blood oxygenation level, parts of the images taken during stimulation should show increased intensity, compared with those taken whilst at rest. The parts of these images which show increased intensity should correspond to the brain areas which are activated by the stimulation.

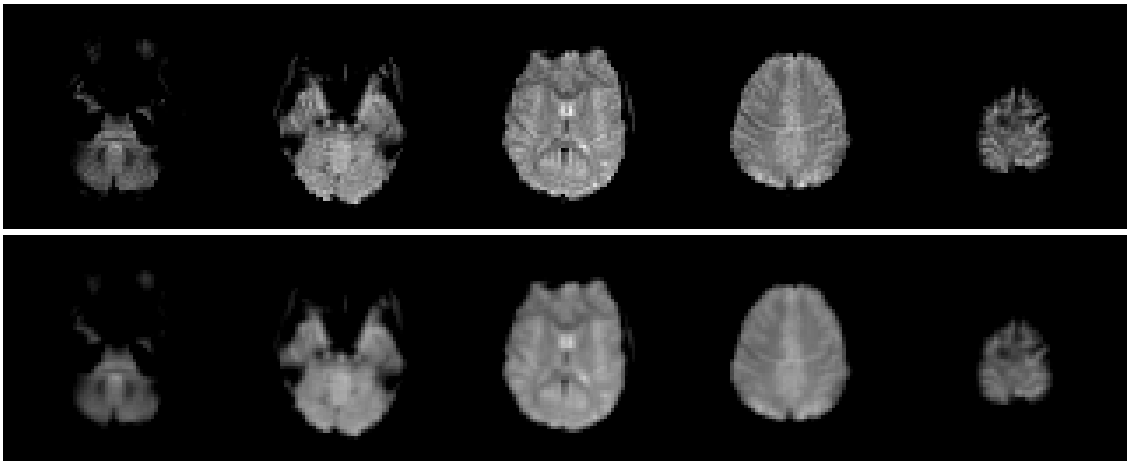

Figure 1. Example slices from an example fMRI volume, before and after spatial filtering of 5mm FWHM.

An example time-series from a single voxel is shown in Figure 2. Image intensity is shown on the y axis, and time (in scans) on the x axis. As described above, for some of the time points, stimulation was applied, (the higher intensity periods), and at some time points the subject was at rest. As well as the effect of the stimulation being clear, the high frequency noise is also apparent. The aim of fMRI analysis is to identify in which voxels' time-series the signal of interest is significantly greater than the noise level.

*For the remainder of this document, reference to "stimulation" should be taken to include also the carrying out of physical or cognitive activity.

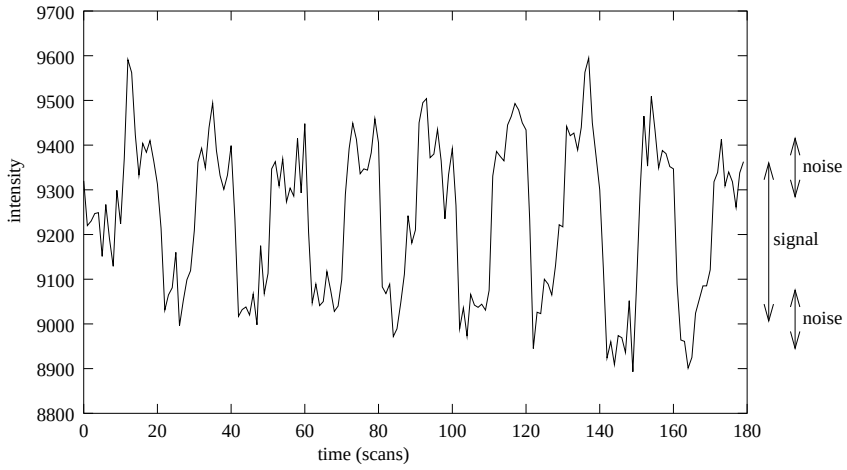

Figure 2. An example time series at a strongly activated voxel from a visual stimulation experiment. Here the signal is significantly larger than the noise level. Periods of stimulation are alternated with periods of rest - a complete stimulation-rest cycle lasts 20 scans.

3. PREPARING FMRI DATA FOR STATISTICAL ANALYSIS

Initially, a 4D data set is pre-processed, i.e., prepared for statistical analysis. Once data has been acquired by the MR scanner, the pre-processing starts by reconstructing the raw "k-space" data into images that actually look like brains. Very often, the next step applied is slice-timing correction; because each slice in each volume is acquired at slightly different times, it necessary to adjust the data so that it appears that all voxels within one volume had been acquired at exactly the same time (all subsequent processing is far simpler if this is done). Each volume is now transformed (using rotation and translation) so that the image of the brain within each volume is aligned with that in every other volume; this is known as motion correction.

Most researchers now blur each volume spatially, principally to reduce noise, hopefully without significantly affecting the activation signal. Figure 1 shows example slices of an FMRI volume before and after a small amount of spatial smoothing. After this, each volume's overall intensity level may be adjusted so that all volumes have the same mean intensity - this intensity normalization can help reduce the effect of global changes in intensity over time. Reduction in low and high frequency noise is normally desired as a final step; each voxel's time series is filtered in order to achieve this.

The purpose of the preprocessing is to remove various kinds of artefacts in the data, and to condition the data, in order to maximize the sensitivity of later statistical analysis, and also, in some situations, to increase the statistical validity.

4. STATISTICAL ANALYSIS OF ACTIVATION IMAGES

In this section we give a very brief overview of different approaches to obtaining activation maps, followed by a slightly more detailed introduction to analysis via the general linear model (GLM – currently the most popular statistical approach) and also various methods of thresholding the resulting statistics maps.

After the pre-processing steps, statistical analysis is carried out to determine which voxels are activated by the stimulation. This can be simple correlation analysis or more advanced modelling of the expected haemodynamic response to the stimulation. Various possible statistical corrections can be included, such as correction for smoothness of the measured time series at each voxel. The main output from this step is a statistical map which indicates those points in the image where the brain has activated in response to the stimulus.

It is most common to analyse each voxel's time series independently ("univariate analysis"). For example, standard general linear model (GLM) analysis is univariate (although cluster-based thresholding, commonly used at the final inference stage, does use spatial neighborhood information and is therefore not univariate). This is the approach which we are going to concentrate on in this document. However, there are also "multivariate" methods (e.g. Friston et al. (1996)) which process all the data together; these methods make more use of spatial relationships within the data than univariate analysis.

There is also a distinction between model-based and model-free methods. In a model-based method, e.g. Friston et al. (1995), a model of the expected response is generated and compared with the data. In a model-free (or data-driven) method, effects or components of interest in the data are found without attempting to use our knowledge of the experimental stimulus, but instead on the basis of a decomposition of the data into uncorrelated time courses or spatial patterns (Singular Value Decomposition - SVD, Principal Component Analysis - PCA), or into statistically independent spatial or temporal components (Independent Component Analysis - ICA) (McKeown et al., 1998; Beckmann and Smith, 2004). This allows us to find unexpected things in the data (either signals of interest or noise artefacts), and is also useful for the analysis of data where it is difficult to generate a good model. Figure 3 shows a comparison of the findings of a model-based approach and a model-free approach for a visual stimulus experiment.

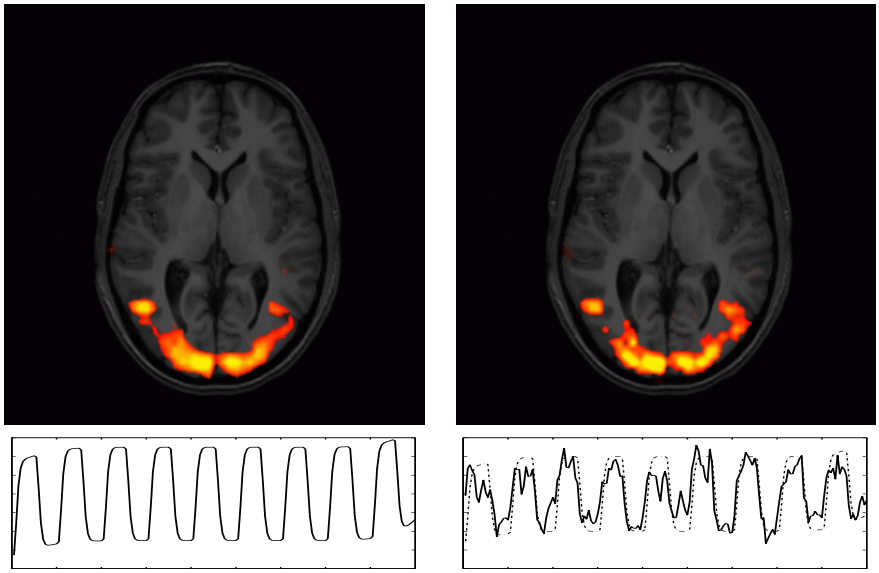

Figure 3. An experiment with alternating periods of rest and visual stimulation. On the left are the areas in the brain labeled "active" using a model-based method, along with the assumed time series modelled as the stimulus convolved with a haemodynamic response function. On the right are the areas in the brain labeled "active" using a model-free method (spatial ICA (Beckmann and Smith, 2004)) for a component whose temporal pattern is also shown. The model-free approach extracts a component corresponding to the visual stimulus without using our knowledge of the stimulus timing. The activation maps are overlaid on a high resolution structural image.

4.1. General Linear Model - Overview

General linear modelling sets up a model (i.e., a general pattern which you expect to see in the data) and fits it to the data. If the model is derived from the timing of the stimulation that was applied to the subject in the MRI scanner, then a good fit between the model and the data means that the data was probably caused by the stimulation.

As the GLM is normally used in a univariate way, the rest of this section considers one voxel only, and the fitting of models to a single voxel's time-course. A very simple example of linear modelling is

$$y(t) = \beta * x(t) + c + e(t), \tag{1}$$

where $y(t)$ is the data, and is a 1D vector of intensity values - one for each time point, i.e., is a function of time. $x(t)$ is the model, and is also a 1D vector with one value for each time point. In the case of a square-wave block design, $x(t)$ might be a series of 1s and 0s - for example, 0 0 0 0 0 1 1 1 1 1 0 0 0 0 0 etc. β is the parameter estimate for $x(t)$, i.e., the value that the square wave (of height 1) must be multiplied by to fit the square wave component in the data. c is a constant, and in this example, would correspond to the baseline (rest) intensity value in the data. e is the error in the model fitting. Thus the model fitting involves adjusting the baseline level and the height of the square wave, to best fit the data; the error term accounts for the residual error between the fitted model and the data.

If there are two types of stimulus, the model would be

$$y = \beta_1 * x_1 + \beta_2 * x_2 + c + e. \tag{2}$$

Thus there are now two different model waveforms corresponding to the two stimulus time-courses. There are also two interesting parameters to estimate, β_1 and β_2. Thus if a particular voxel responds strongly to model x_1 the model-fitting will find a large value for β_1; if the data instead looks more like the second model time-course, x_2, then the model-fitting will give β_2 a large value. Different model waveforms within a complex model are often referred to as explanatory variables (EVs), as they explain different processes in the data.

4.1.1. Haemodynamic response function

In order to get the best possible fit of the model to the data, the "stimulus function" (which is often a sharp on/off waveform) is convolved with the haemodynamic response function (HRF) (Friston et al., 1994). This process mimics the effect that the brain's neuro-physiology has on the input function (the stimulation). The brain's haemodynamic response is a delayed and blurred version of the input time-series, so a mathematical operation is applied to the stimulus function to take the square wave input and create a delayed and blurred version, which will better fit the data. For example, see Figure 4, showing the raw stimulation timing waveform and the HRF-convolved model - $x(t)$ - which will be used in the model fitting. It is well established that the HRF varies between brain regions and subjects (Josephs et al., 1997). Flexibility in the HRF can be achieved within the GLM by the use of basis sets (Josephs et al., 1997). Different linear combinations of the basis sets can give a range of HRF shapes. A popular basis set is the stimulus convolved with a canonical HRF shape along with its temporal derivative (Friston et al., 1998a). This can account for shifts in time of the stimulus convolved HRF.

It is important to appreciate that non-linearities are present when there are short separations (less than approximately 3 seconds) between stimuli (Friston et al., 1998b). These experimental designs should be avoided, or nonlinearities need to be modelled using techniques such as biophysical modelling (Buxton et al., 1998) or Volterra kernels (Friston et al., 1998b).

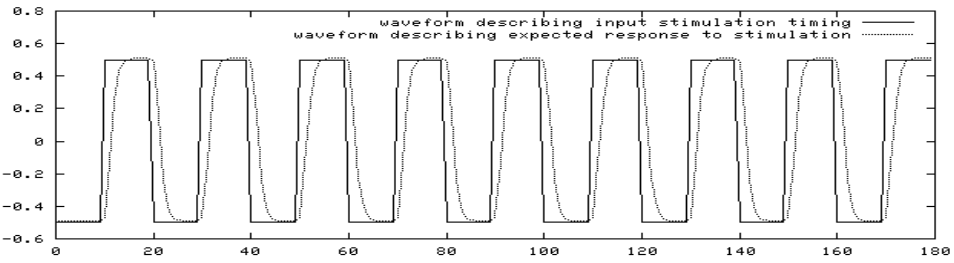

Figure 4. Model waveform formation: the square waveform describes the input stimulation timing; the smoothed waveform results from convolving the first with the haemodynamic response function, a transformation which leaves the model looking much more like the measured data.

4.1.2. Design Matrix

The GLM is often formulated in matrix notation. Thus all of the parameters are grouped together into a vector β, and all of the model time-courses are grouped together into a matrix $\mathbf{X}$, often referred to as the design matrix. Figure 5 shows an example design matrix with two such model time-courses. Each column is a different part of the model. For example, in a particular experiment, both visual and auditory stimulations are applied, but with different timings; the left column (x_1 or EV 1) models the visual stimulation, and the right column (EV 2 or x_2) models the auditory stimulation.

4.1.3. Parameter estimation and temporal autocorrelation

When the model is fit separately to the data at each voxel, there will be found an estimate of the "goodness of fit", of each column in the model, to that voxel's time-course. In the example visual/auditory experiment, the first column will generate a high first parameter estimate in the visual cortex. However, the second column will generate a low second parameter estimate, as this part of the model will not fit the voxel's time-course well. Mathematically, these parameter estimates are obtained by using least squared estimators (they minimize the squared difference between the model fit and the data).

Ideally, all of the EVs should be independent of each other. If any EV is close to being a sum (or weighted sum) of other EVs in the design, then the fitting of the model to the data does not work properly (the design matrix is not of "full

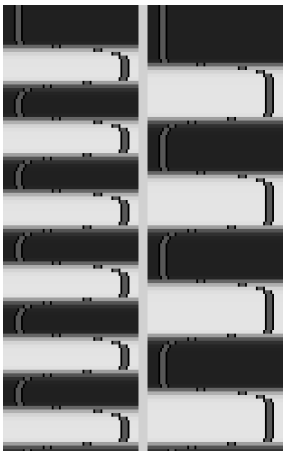

Figure 5. Example design matrix with two explanatory variables; two different stimulations are being applied. Because they have different timings, they are modelled separately - see Equation 2. Time is on the y axis, pointing downwards. Note that in this particular visual representation of a design matrix, each column has two representations of the model's value at each point in time: the underlying intensity encodes the model's value at a particular time point, and so does the line graph.

rank"). A common mistake of this type is to model both rest and activation waveforms, making one an upside-down version of the other; in this case EV 2 is -1 times EV 1, and therefore linearly dependent on it. It is only necessary to model the activation waveform.

The presence of temporal autocorrelation in FMRI noise affects the best estimation to use (and also importantly the accuracy of subsequent statistical tests). Typically, this requires the parameter estimation to be combined with some form of temporal autocorrelation estimation and temporal filtering. The most commonly used approach is "prewhitening", which uses a temporal filter to effectively remove the autocorrelation (Bullmore et al., 1996; Friston et al., 2000; Woolrich et al., 2001).

A wide range of different strategies have been proposed for dealing with temporal autocorrelation in FMRI. These include autoregressive (AR) models (Bullmore et al., 1996; Locascio et al., 1997), AR plus white noise (Purdon and Weisskoff, 1998), spatial regularisation of autocorrelation estimates (Woolrich et al., 2001; Worsley et al., 2002; Gautama and Van Hulle, 2004), and wavelet resampling (Bullmore et al., 2001).

4.1.4. T-statistics

To convert a parameter estimate (PE, i.e. the estimated β value) into a useful statistic, its value is compared with the uncertainty in its estimation (resulting in what is known as a T-statistic or T-value; $T = PE/\text{standard error}(PE)$). If the PE is low relative to its estimated uncertainty, the T-value will be low, and the fit is less likely to be significant (and vice versa). (Remember, all of this is carried out separately for each voxel.)

The question remains as to how we determine that a T-value corresponds to a significant fit. This is achieved by comparing the calculated T-value to the distribution of T-values we would expect to get if the true β value was zero. This is a **null hypothesis test** (the null hypothesis is that β is zero). This gives the probability, or P-value, of getting a T-value greater than the one we have calculated if the null hypothesis were true. A low probability (low P-value) of the null hypothesis being true means that we can more confidently reject the null hypothesis and label the voxel as having a significant fit (later we will see how we choose a threshold value for the P-values).

Typically, we can assume that the noise in FMRI is Gaussian distributed. This means that the expected distribution of T-values under the null hypothesis is T-distributed and a P-value can be easily calculated. Note that we can convert a T-value into a P-value, and then into a Z-value/statistic [†] using standard statistical transformations; however, T, P and Z all contain the same information - they describe how significantly the data is related to a particular part of the model (x_1 or x_2).

[†] Z is a "Gaussianised T"; if the noise in the data is Gaussian, then T can be simply converted (taking into account the number of time points) into Z. When there is no activation, Z follows a Gaussian distribution, with zero mean and unit variance.

4.1.5. Contrasts

As well as producing images of Z values which describe how strongly each voxel is related to each EV (one image per EV), parameter estimates can be compared to test directly whether one EV is more "relevant" to the data than another. To do this, one PE is subtracted from another, the standard error for this new value is calculated, and a new T image is created. All of the above is controlled by setting up "contrasts". To compare two EVs (for example, to subtract one stimulus type - EV 1 - from another type - EV 2), set EV 1's contrast value to -1 and EV 2's to 1. This is often written as a contrast of [-1 1], as the contrasted parameter estimate equals $-1 \times \beta_1 + 1 \times \beta_2$. A T statistic image will then be generated according to this request, answering the question "where is the response to stimulus 2 significantly greater than the response to stimulus 1?"

It is possible that the response to two different stimuli, when applied simultaneously, is greater than that predicted by adding up the responses to the stimuli when applied separately. If this is the case, then such "nonlinear interactions" need to be allowed for in the model. The simplest way of doing this is to setup the two originals EVs, and then add an interaction EV, which will only be "up" when both of the original EVs are "up", and "down" otherwise. In Figure 6, EV 1 could represent the application of a drug, and EV 2 could represent visual stimulation. EV 3 will model the extent to

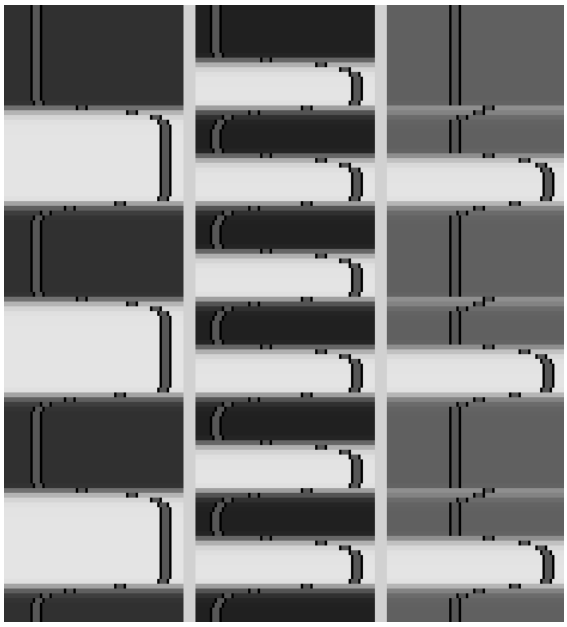

Figure 6. Example of modelling a nonlinear interaction between stimuli. The first two EVs model the separate stimuli, whilst the third models the interaction, i.e., accounts for the "extra" response when both stimuli are applied together.

which the response to drug&visual is greater than the sum of drug-only and visual-only. A contrast of [0 0 1] will show this measure, whilst a contrast of [0 0 -1] shows where negative interaction is occurring.

With "parametric designs", there are typically several different levels of stimulation, and it is common to estimate the response to each level separately. This requires a separate EV for each stimulation level. The different contrasts ask different questions about these responses: For example, for an experiment with 3 stimulation levels, [1 0 0] shows the response to stimulation level 1 versus rest (likewise [0 1 0] for level 2 vs rest and [0 0 1] for level 3). [-1 1 0] shows where the response to level 2 is greater than that for level 1. [-1 0 1] shows the general linear increase across all three levels. [1 -2 1] shows where the level dependence deviates, in an upwardly-curving way, from being linear.

4.2. Inference ("Thresholding")

Thus we now have a statistic map (for example, T, P or Z). The next step is to threshold this, in order to decide, at a given level of significance, which parts of the brain were activated. There are a variety of ways of carrying out thresholding. These are now briefly outlined.

Recall that the P-value at a voxel corresponds to the probability of getting a T-value greater than the one we have calculated if the null hypothesis were true. The simplest method of thresholding is to select a significance threshold for

the P-value and apply this to every voxel in the statistic map. The chosen P-value threshold can be interpreted as the probability of wrongly rejecting the null hypothesis for each null hypothesis test carried out, i.e. the rate at which we expect false positives to occur.

A problem with this is that there are many tests being carried out, because there are so many voxels in the brain. For example, if 20000 voxels are tested for at a significance of $P < 0.01$ (a false positive rate of 0.01) then it is expected that 200 will be called significant by chance, even if no stimulation is applied (200 false positives). It is not ideal to blindly accept these as being activated! This "multiple-comparison problem" means that it is not valid to accept all activations reported by this method of thresholding; a correction is necessary to reduce the number of false positives. Typically a Bonferroni correction is used, where the significance level at each voxel is divided by the number of voxels; this corrects for the number of comparisons being made. However, this results in very stringent thresholding (i.e. in the case given above, the resulting P threshold is $0.01/20000 = 0.0000005$).

A refinement of the above voxel-wise thresholding is to use Gaussian random field (GRF) theory to threshold the image Worsley et al. (1996). The main difference is that this method takes into account the spatial smoothness of the statistic map (i.e. the statistical tests at each voxel are not independent). This method is less "over-conservative" than simple voxel-wise thresholding with Bonferroni correction; typically the correction to P values is reduced (compared with Bonferroni correction) by a factor of 2-20.

Finally, it is possible, again using GRF theory, to take into account spatial extent (i.e. size) of clusters of activations, before estimating significance (Friston et al., 1994). Thus instead of assigning a P-value to each voxel, clusters of voxels are created on the basis of an initial thresholding, and then each cluster is assigned a P-value, which may or may not pass the final significance test. It is often the case that this method is more sensitive to activation than the voxel-based methods. A limitation is the arbitrary nature of the initial thresholding, used to create the clusters.

These approaches control the false positive rate over multiple tests (the expected proportion of all tests that are false positives). An alternative is the False Discovery Rate (FDR), which instead looks to control the expected proportion of *rejected* tests that are false positives (Genovese et al., 2002; Nichols and Hayasaka, 2003). FDR controlling procedures are more powerful, yet still control false positives in a useful manner.

5. MULTI-SUBJECT STATISTICS

Although so far we have only discussed single-session analyses, it is common to run an experiment several times, either on the same subject, or with several different subjects, or both. This can both increase the sensitivity of the overall experiment (as more data can lead to increased sensitivity to an effect) or allow the generalization of any conclusions to the whole population.

In order to combine statistics across different sessions or subjects, the first necessary step is to align the brain images from all sessions into some common space. This is typically done using generic registration tools and can be carried out either on the raw data (i.e. before the single-session, or "first-level", analyses) or on the statistic maps created by the first-level analyses.

Once all the data is aligned, there are a variety of statistical methods for combining results across sessions or subjects, to either create a single result for a group of subjects, or to compare different groups of subjects (for example, placebo group versus drug treatment group). These methods include "fixed-effects" and "mixed-effects" analyses. Fixed-effects assumes that all subjects activate equally, and is only interested in within-session errors. Mixed-effects analysis additionally takes into account between-session errors, and therefore makes less assumptions about the data; its results are therefore valid for the whole population from which the group of subjects is drawn (whereas fixed-effects results are not). However, the mixed-effects analysis tends to give more "conservative" results. For more detail on group modelling and estimation see Friston et al. (2002); Beckmann et al. (2003); Woolrich et al. (2004a).

Figure 7 shows the results of the significant differences between the two subject groups in a pain-warm contrast.

6. REGISTRATION AND BRAIN ATLASES

As described above, registration (aligning different brain images) is often used when combining fMRI data from different sessions or subjects. It can also be used to align the low-resolution fMRI images to a high-resolution structural image,

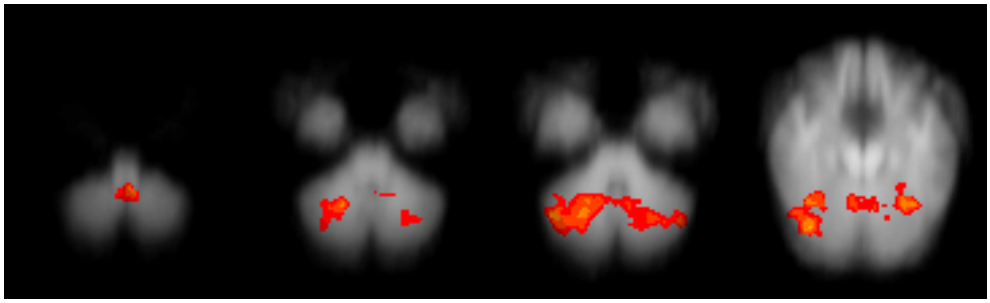

Figure 7. Significant differences between the two subject groups in a pain-warm contrast.

so that the activations can be viewed in the context of a good quality brain image; this aids in the interpretation of the activations.

A related issue is the use of templates and brain atlases. These consist of data which is transformed into some "standard brain space", for example the co-ordinate system specified by Talairach and Tournoux (1988). A template is typically an average of many brains, all registered into any given common co-ordinate system. An example is the MNI 305 average (Collins et al., 1994). An atlas is also based in a common co-ordinate system, but contains more sophisticated information about the brain at each voxel, for example, information about tissue type, local brain structure or functional area. Atlases can inform interpretation of fMRI experiments in a variety of ways, helping the experimenter gain the maximum value from the data.

7. ALTERNATIVE METHODS

Most of the methods we have described for inferring brain activity from FMRI are based on using parametric statistics (e.g. we assume the FMRI noise is Gaussian distributed). However, nonparametric approaches using permutation methods have also been developed for null hypothesis testing. These have the advantage of being more robust to assumptions, and can deal with tests intractable to parametric statistics (Bullmore et al., 1999; Nichols and Holmes, 2001). The main disadvantage is that these techniques tend to be less sensitive than parametric approaches.

An emerging alternative to frequentist null hypothesis testing is to use Bayesian inference. This provides the framework for a wider range of inference questions to be asked, for inferring on a wide range of models, and for incorporating probabilistic prior information. For example, Bayesian methods have been used with nonlinear HRF models (Genovese, 2000; Gössl et al., 2001; Friston, 2002; Marrelec et al., 2003; Woolrich et al., 2004c), hierarchical multi-subject/session FMRI models (Friston et al., 2002; Woolrich et al., 2004a), and adaptive spatial regularisation of FMRI model parameters (Woolrich et al., 2004b; Penny et al., 2005).

References

Beckmann, C., Jenkinson, M., and Smith, S. (2003). General multi-level linear modelling for group analysis in FMRI. *NeuroImage*, 20:1052–1063.

Beckmann, C. and Smith, S. M. (2004). Probabilistic Independent Component Analysis for Functional Magnetic Resonance Imaging. *IEEE Trans. on Medical Imaging*, 23(2):137–152.

Bullmore, E., Brammer, M., Williams, S., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R., and Sham, P. (1996). Statistical methods of estimation and inference for functional MR image analysis. *Magnetic Resonance in Medicine*, 35(2):261–277.

Bullmore, E., Long, C., Suckling, J., Fadili, J., Calvert, G., Zelaya, F., Carpenter, T. A., and Brammer, M. (2001). Colored noise and computational inference in neurophysiological (fMRI) time series analysis: resampling methods in time and wavelet domains. *Hum Brain Mapp*, 12(2):61–78.

Bullmore, E., Rabe-Hesketh, S., Morris, R., Williams, S., Gregory, L., Gray, J., and Brammer, M. (1999). Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural mr images of the brain. *IEEE Trans. on Medical Imaging*, 418:32–42.

Buxton, R., Wong, E., and Frank, L. (1998). Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magnetic Resonance in Medicine*, 39:855–864.

Collins, D., Neelin, P., Peters, T., and Evans, A. (1994). Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *Journal of Computer Assisted Tomography*, 18(2):192–205.

Friston, K., Fletcher, P., Josephs, O., Holmes, A., Rugg, M., and Turner, R. (1998a). Event-related fMRI: Characterizing differential responses. *NeuroImage*, 7:30–40.

Friston, K., Holmes, A., Worsley, K., Poline, J.-B., Frith, C., and Frackowiak, R. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2:189–210.

Friston, K., Josephs, O., Rees, G., and Turner, R. (1998b). Nonlinear event-related responses in fMRI. *Magnetic Resonance in Medicine*, 39:41–52.

Friston, K., Josephs, O., Zarahn, E., Holmes, A., Rouquette, S., and Poline, J.-B. (2000). To smooth or not to smooth? *NeuroImage*, 12:196–208.

Friston, K., Poline, J.-B., Holmes, A., Frith, C., and Frackowiak, R. (1996). A multivariate analysis of PET activation studies. *Human Brain Mapping*, 4:140–151.

Friston, K., Worsley, K., Frackowiak, R., Mazziotta, J., and Evans, A. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1:214–220.

Friston, K. J. (2002). Bayesian estimation of dynamical systems: An application to fMRI. *NeuroImage*, 16:513–530.

Friston, K. J., Glaser, D. E., Henson, R. N. A., Kiebel, S., Phillips, C., and Ashburner, J. (2002). Classical and Bayesian inference in neuroimaging: Applications. *NeuroImage*, 16:484–512.

Gautama, T. and Van Hulle, M. M. (2004). Optimal spatial regularisation of autocorrelation estimates in fMRI analysis. *Neuroimage*, 23(3):1203–1216.

Genovese, C. (2000). A Bayesian time-course model for functional magnetic resonance imaging data (with discussion). *Journal of the American Statistical Association*, 95:691–703.

Genovese, C. R., Lazar, N. A., and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4):870–878.

Gössl, C., Auer, D., and Fahrmeir, L. (2001). Bayesian modeling of the haemodynamic response function in BOLD fMRI. *NeuroImage*, 14(1):140–148.

Jezzard, P., Matthews, P., and Smith, S., editors (2001). *Functional MRI: An Introduction to Methods*. OUP, Oxford.

Josephs, O., Turner, R., and Friston, K. (1997). Event-related fMRI. *Human Brain Mapping*, 5:1–7.

Locascio, J., Jennings, P., Moore, C., and Corkin, S. (1997). Time series analysis in the time domain and resampling methods for studies of functional magnetic resonance brain imaging. *Human Brain Mapping*, 5:168–193.

Marrelec, G., Benali, H., Ciuciu, P., Pélégrini-Issac, M., and Poline, J.-B. (2003). Robust Bayesian estimation of the hemodynamic response function in event-related BOLD MRI using basic physiological information. *Hum. Brain Mapp.*, 19(1):1–17.

McKeown, M. J., Makeig, S., Brown, G. G., Jung, T. P., Kindermann, S. S., Bell, A. J., and Sejnowski, T. J. (1998). Analysis of fMRI data by blind separation into independent spatial components. *Human Brain Mapping*, 6(3):160–88.

Nichols, T. and Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Stat Methods Med Res*, 12(5):419–446.

Nichols, T. E. and Holmes, A. P. (2001). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15:1–25.

Penny, W., Trujillo-Barreto, N., and Friston, K. (2005). Bayesian fMRI time series analysis with spatial priors. *NeuroImage*, 24(2):350–362.

Purdon, P. and Weisskoff, R. (1998). Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates in fMRI. *Human Brain Mapping*, 6:239–249.

Talairach, J. and Tournoux, P. (1988). *Co-planar Stereotaxic Atlas of the Human Brain.* Thieme Medical Publisher Inc., New York.

Woolrich, M., Behrens, T., Beckmann, C., Jenkinson, M., and Smith, S. (2004a). Multi-level linear modelling for FMRI group analysis using Bayesian inference. *NeuroImage*, 21(4):1732–1747.

Woolrich, M., Behrens, T., and Smith, S. (2004b). Constrained linear basis sets for HRF modelling using Variational Bayes. *NeuroImage*, 21(4):1748–1761.

Woolrich, M., Jenkinson, M., Brady, J., and Smith, S. (2004c). Fully Bayesian spatio-temporal modelling of FMRI data. *IEEE Trans. on Medical Imaging*, 23(2):213–231.

Woolrich, M., Ripley, B., Brady, J., and Smith, S. (2001). Temporal autocorrelation in univariate linear modelling of FMRI data. *NeuroImage*, 14(6):1370–1386.

Worsley, K., Liao, C., Aston, J., Petre, V., Duncan, G., Morales, F., and Evans, A. (2002). A general statistical analysis for fMRI data. *NeuroImage*, 15(1):1–15.

Worsley, K., Marrett, S., Neelin, P., Vandal, A., Friston, K., and Evans, A. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 4:58–73.